A STATISTICAL INVESTIGATION ON MACHINE LEARNING BASED MODELLING OF DIABETES MELLITUS

Sanmati Kumar Jain [1], Dr. Nidhi Mishra [2]

[1] Ph.D. Scholar, Department of Computer Science and Engineering, Kalinga University, Raipur, CG, India

[2] Professor, Department of Computer Science and Engineering, Kalinga University, Raipur, CG, India

Abstract

Diabetes Mellitus (DM) is a chronic metabolic disorder that poses significant global health challenges, particularly with the rising incidence of Type 2 Diabetes Mellitus (T2DM). This study presents a machine learning-based predictive framework for early diagnosis of T2DM using lifestyle and biological data collected from a diverse population. After data collection and preprocessing, class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Multiple machine learning models, including Bagged Decision Trees (BDT), Stochastic Gradient Boosting (SGB), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Linear Regressor, were developed and compared. Among them, BDT achieved the highest performance with 99.05% accuracy and 99.56% recall. The study demonstrates the effectiveness of ensemble learning techniques in enhancing predictive accuracy and reducing false negatives in medical diagnostics. The results suggest that integrating machine learning models with practical health data can significantly improve early detection and management of T2DM, especially in resource-constrained healthcare environments.

Keywords: Diabetes, Machine Learning, Predictive Modeling, Healthcare, Diagnosis

1. Introduction

Diabetes mellitus is a group of metabolic disorders characterized by elevated blood glucose levels due to inadequate insulin production or improper insulin response in the human body. There are around 40 forms of diabetes, and individuals worldwide are mostly unaware of the complications associated with this condition due to an under-resourced healthcare system. Several prevalent categories include [3]: Type 1 Diabetes Mellitus is reliant on insulin and

medical intervention and may manifest at any age, particularly in children [4]. Type 2 Diabetes Mellitus is characterized by insulin independence and is influenced by lifestyle factors; it is the most prevalent form of diabetes affecting individuals across all age demographics. [5]. Gestational diabetes is induced by hyperglycemic circumstances, occurring in pregnant women and potentially impacting newborns. Pre-diabetes is attributed to elevated insulin production resulting from genetic abnormalities [7]. Diabetes may directly impact several organs of the human body, including the kidneys, brain, and liver. The prevalent first symptoms and indicators seen in diabetic and possibly diabetic individuals are polydipsia, profound weariness, skin discoloration, and polyuria, among others. Numerous significant statistical data from different health organizations on diabetes mellitus indicate the danger of developing life-threatening, severe, and catastrophic complications [9,10]. The International Diabetes Federation (IDF) Atlas 2019 reports that 463 million persons aged 20 to 79 years, constituting 9.3% of the global population, are now living with diabetes. Projections indicate that it will impact 578 million individuals by 2030 and 700 million by 2045 [11]. In 2019, it was predicted that diabetes mellitus caused 4.2 million deaths globally. According to IDF Atlas 2019, it is projected that by 2045, India would lead the world with an estimated [134.3–165.2] million individuals with diabetes [2]. Figure 1 illustrates the seven nations or areas most affected by diabetes, together with the undiagnosed population in millions [2].

Recent research indicates that Machine Learning (ML) and Ensemble Learning (EL) approaches significantly contribute to the prediction of many chronic illnesses, such as diabetes, yielding favorable outcomes based on several statistical metrics [12–15]. Early prediction of diabetes mellitus is crucial, and achieving a greater accuracy rate by machine learning approaches is imperative [16,17]. The primary aim of this project is to develop a computational model using machine learning and ensemble learning methods with lifestyle data to predict Type 2 Diabetes Mellitus (TIIDM). This study employs Bagged Decision Tree (BDT) and Stochastic Gradient Boosting (SGB) using the Boosting technique.

2. Data sampling and methodology

2.1. Dataset handling

The dataset used for this research was gathered using both offline and online methods, following the guidance of domain experts. The lifestyle factors were established in collaboration with diabetes specialists, including endocrinologists and diabetologists. The authors actively participated in data collection for this study during a two-year period. Survey forms have been sent to various hospitals to gather data from distinct departments such as inpatient, outpatient, and emergency, categorized by different demographic areas. Google Forms were created to collect data on individuals employed in diverse organizations, ensuring a comprehensive representation of numerous categories. It comprises individuals from various locations (both rural and urban), adults across diverse age demographics, and a balanced male-to-female ratio. The dataset consisted of 1,939 records and 11 biological/lifestyle characteristics, with the first 10 parameters serving as independent variables and the last parameter (Outcome) as the dependent variable.

2.2. Data pre-processing

Data pre-processing is essential for achieving improved outcomes prior to using distinct statistical libraries inside the Integrated Development Environment Spyder, employing Python (3.9.1) as the programming language [26,27].

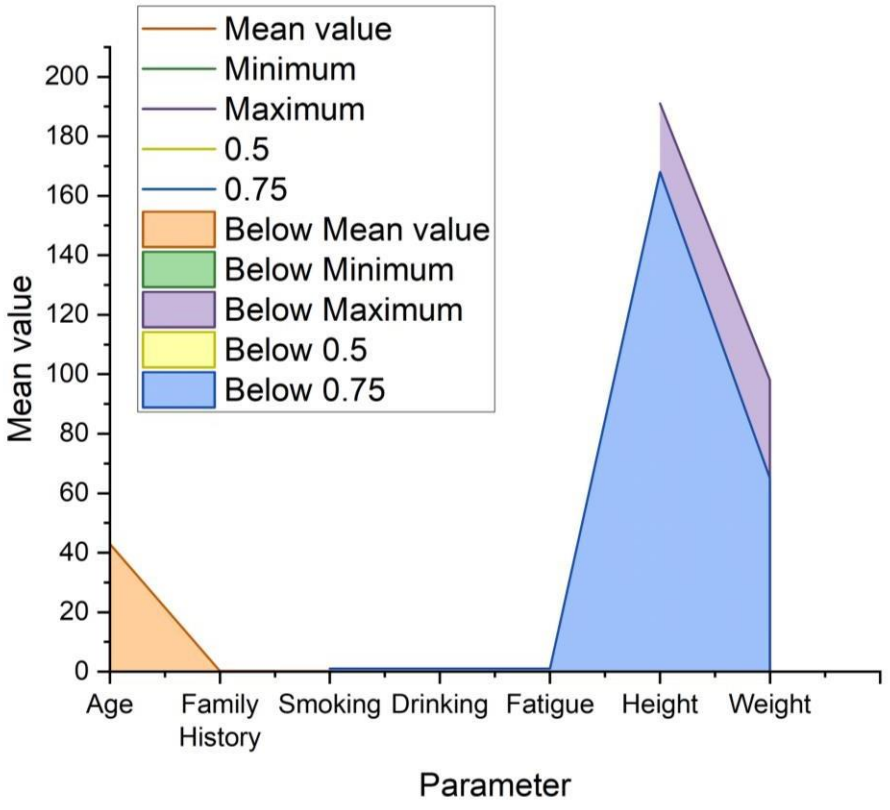

Figure 1 Plot representation of sample distribution values

The necessary libraries for data quality evaluation have been imported, and missing values have been addressed using the imputation approach by averaging certain attribute values such as Age, Sex, Height, Weight, Thirst, and Fatigue. Outlier identification was conducted using a boxplot, using the interquartile range approach to substitute outliers with viable sample values. Data transformation has been performed to enhance the efficiency of data prior to constructing the machine learning models. Furthermore, duplication, inconsistency, and damaged data have been eradicated from the dataset by the use of several exploratory data analysis approaches [28].

Table 1 Data description used for study

Parameter	Age (in years)	Family History	Smoking	Drinking	Fatigue	Height (cm)	Weight (Kg)
Mean value	42.98	0.24	0.18	0.22	0.65	161	64
Minimum	7	-	-	-	-	58	14

Maximum	89	-	-	-	-	191	98
50%	38	-	-	-	1	161	59
75%	52	-	1	1	1	168	65

2.3. SMOTE for data balancing

The Synthetic Minority Oversampling Technique (SMOTE) is an effective method often used for addressing class imbalance in high-dimensional data to resolve many real-world issues [29]. The primary challenge when dealing with unbalanced datasets is that model development often results in suboptimal performance across multiple statistical metrics. This research used the SMOTE approach for class balancing prior to the development of the ML/EL models to enhance the predictive capabilities of the framework. This approach facilitates the oversampling and augmentation of the minority class within the dataset. It randomly picks instances from the minority class, identifies their k closest minority class neighbors, and then selects one neighbor to create a line segment in the feature space using a convex combination of two synthetic examples, referred to as A and B. Figure 3 illustrates the count of the Outcome (class variable) before to and after to the use of the SMOTE method [30].

3. ML models and ensembling

Machine learning approaches, when used, have yielded superior outcomes across almost all domains, particularly in healthcare analytics [32–35]. Machine learning models are being strategically integrated to enhance the analytical capabilities of various created frameworks to address real-world issues. In Bagging, several models with similar learning characteristics are generated from distinct subsamples of the training set. Conversely, the Boosting approach involves the development of several models of the same kind, each designed to rectify the prediction mistakes of its predecessor in a sequential manner. Ultimately, several models of distinct types have been constructed in the Voting technique, and their predictions have been aggregated by calculating the mean to enhance the diverse statistics and machine learning metrics for improved prediction of TIIDM illness.

3.1. Bagging method

Bagging is an ensemble strategy that utilizes the bootstrap aggregation method to generate numerous training sets for model development purposes. The training sets are generated from the original dataset in a randomly repeated manner. Following the development of diverse training sets, several models are used in the resampling process using an ensemble framework. Ultimately, the outcomes of the learners are consolidated to get the final forecast. It aids in diminishing the variations from the models during training to address the issue of overfitting. The three primary phases used in the bagging approach are bootstrapping, parallel training, and aggregation.

3.2. Boosting method

Boosting is a method of transforming weak learners into strong learners. This strategy amalgamates all conventional or weak classifiers to create a robust model, hence enhancing the predicted accuracy of the final outcomes.

3.3. Support Vector Machine (SVM)

Support Vector Machine is a powerful supervised learning algorithm used for both classification and regression tasks. In regression, SVM tries to find a function that deviates from actual target values by a value no greater than a specified margin (epsilon) while maintaining model simplicity. It works well in high-dimensional spaces and can handle non-linear relationships using kernel functions like radial basis function (RBF) or polynomial kernels. SVM is particularly useful when the dataset has clear margins of separation or when avoiding overfitting is a priority.

3.4. Multilayer Perceptron (MLP)

Multilayer Perceptron is a class of feedforward artificial neural networks consisting of an input layer, one or more hidden layers, and an output layer. Each layer is made up of interconnected neurons that use non-linear activation functions such as ReLU or Tanh to learn complex relationships in the data. MLP is trained using backpropagation and gradient descent algorithms. It is effective in capturing non-linear patterns and is widely used in applications like classification, regression, and time-series prediction. Due to its flexibility and learning capacity, MLP is well-suited for modeling complex engine behaviors.

3.5. Linear Regressor

Linear regression is one of the simplest and most interpretable machine learning algorithms. It assumes a linear relationship between the input features and the output variable. The model fits a straight line (in multiple dimensions, a hyperplane) by minimizing the difference between actual and predicted values using a cost function, typically mean squared error (MSE). Although linear regression may not capture complex patterns, it performs well for datasets with linear trends and serves as a useful baseline for regression tasks.

4. Results and discussion

This work utilized two ensemble-based machine learning models—Bagged Decision Trees (BDT) and Stochastic Gradient Boosting (SGB)—for the early prediction of Type II Diabetes Mellitus (TIIDM) using lifestyle and biological factors gathered from a heterogeneous population. The dataset, pre-processed and balanced by SMOTE, effectively mitigated class imbalance, a significant issue in medical data analytics. Table 2 displays the performance of the ML models using two essential assessment metrics: Accuracy and Recall. Both ensemble models attained remarkable performance, with BDT achieving an accuracy of 99.05% and a recall of 99.56%, slightly outperforming SGB, which achieved 97.79% accuracy and 98.06% recall.

In addition to the ensemble methods, three standalone machine learning models—Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Linear Regressor—were also evaluated. SVM demonstrated a strong performance with an accuracy of 97.81% and a recall of 97.93%, while MLP achieved 98.06% accuracy and 97.62% recall, reflecting its ability to capture non-linear patterns. The Linear Regressor, though simpler in nature, achieved a respectable accuracy of 94.68% and a recall of 97.04%, making it a useful baseline model.

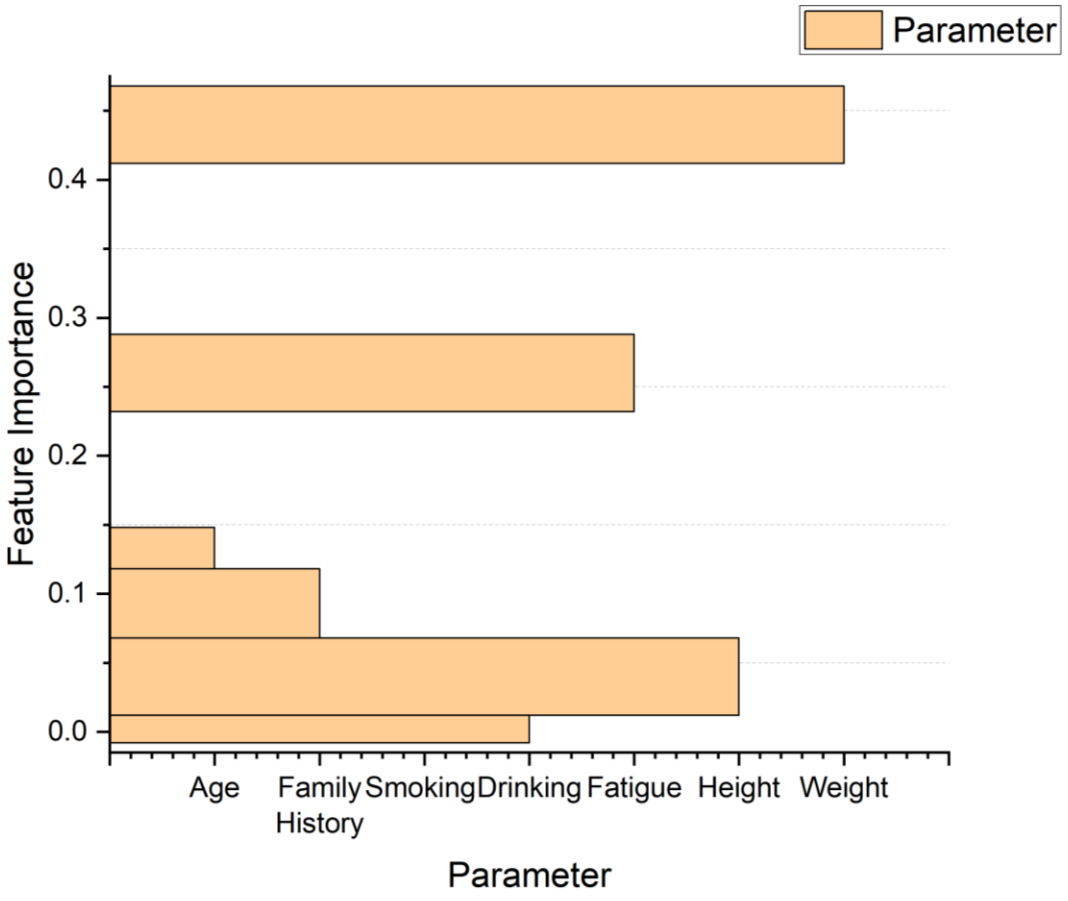

Figure 2 Feature Importance

The findings demonstrate that ensemble models are particularly proficient at identifying nuanced patterns and correlations among predictor variables. The marginal superiority of BDT in recall indicates its effectiveness in detecting true positive cases (actual diabetes patients), which is vital for minimizing false negatives in a clinical setting. This ability to reduce diagnostic errors, especially undiagnosed cases, reinforces the value of ensemble techniques in healthcare analytics. Bagging effectively reduces variance by combining multiple weak learners trained on different data subsets, thereby preventing overfitting. Boosting enhances model learning by sequentially correcting previous errors. These ensemble techniques, alongside high-performing standalone models, are especially suitable for complex and noisy datasets often found in healthcare.

Moreover, the inclusion of practical lifestyle features—such as fatigue, thirst, smoking habits,

and family history—enhances model interpretability and aligns with clinical insights. This supports the real-world applicability of these models in early diabetes screening, particularly in resource-limited healthcare systems.

These results highlight the promise of machine learning-based screening techniques in resource-constrained settings. High-performing models may aid physicians by functioning as decision-support tools, alleviating diagnostic burden, and enhancing early detection rates. In nations such as India, where the anticipated diabetes population is very large, these prediction algorithms may significantly contribute to preventive healthcare planning.

Table 2 Classification performance measurements of test dataset

Algorithm	Accuracy	Recall
Bagged decision trees	99.05	99.56
Stochastic Gradient Boosting	97.79	98.06
SVM	97.81	97.93
MLP	98.06	97.62
Linear Regressor	94.68	97.04

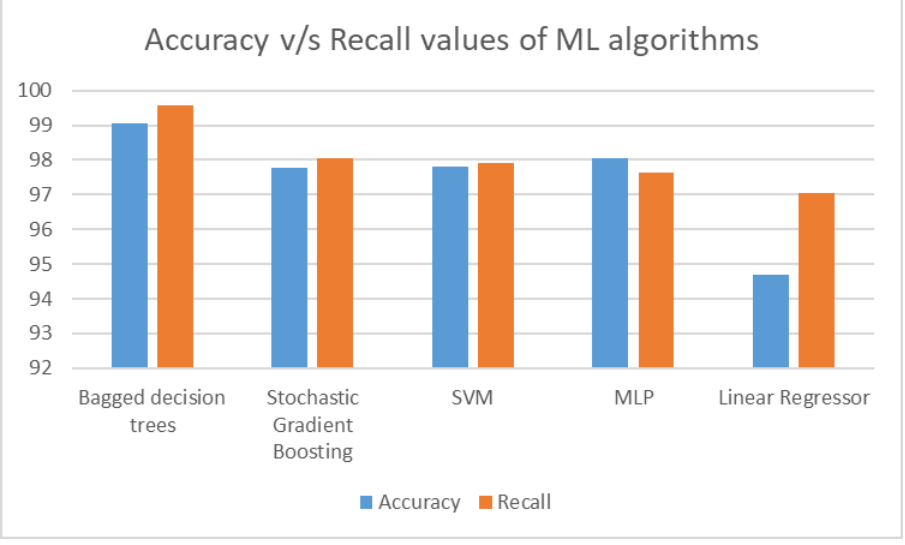

Figure 3 Graphical representation of performance measurements

5. Conclusion

This research highlights the successful application of machine learning and ensemble models in predicting Type 2 Diabetes Mellitus using lifestyle and biological indicators. The use of SMOTE ensured balanced datasets, enhancing model reliability. Among the tested models, Bagged Decision Trees showed the highest performance in both accuracy and recall, proving effective in minimizing false negatives—a critical factor in medical diagnosis. The inclusion of real-world features such as fatigue, thirst, smoking habits, and family history contributed to the model's interpretability and clinical relevance. These results affirm the potential of AI-driven systems as decision-support tools for early diabetes screening. Especially in high-risk regions like India, such tools can play a pivotal role in preventive healthcare, alleviating diagnostic burdens and enabling timely medical intervention.

.

References

[1] T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, J. Big Data 6 (1) (2019) http://dx.doi.org/10.1186/ s40537-019-0175-6.

[2] Diabetes Federation International, IDF, DF Diabetes Atlas 2019, ninth ed., 2019.

[3] M. Sharma, Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: A review, Int. J. Pharm. Sci. Res. 9 (7) (2018) 2700–2719.

[4] R. Sengamuthu, R. Abirami, D. Karthik, Various data mining techniques analysis to predict, Int. Res. J. Eng. Technol. 5 (5) (2018) 676–679, [Online]. Available: https://www.irjet.net/archives/V5/i5/IRJET-V5I5134.pdf.

[5] D. J. D. M.D., Diabetes mellitus diabetes mellitus, Ferri's Clin. Advis. 2020, 512 (January) (2020) 432–441.

[6] An, D. Shakti, Prediction of diabetes based on personal lifestyle indicators, in: Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol., no. September, NGCT 2015, 2016, pp. 673–676.

[7] S. Vyas, R. Ranjan, N. Singh, A. Mathur, Review of predictive analysis techniques for analysis diabetes risk, in: Proc. - 2019 Amity Int. Conf. Artif. Intell., AICAI 2019, 2019, pp. 627–631.

[8] D.M. Chan, Director-General, WHO, Global Report on Diabetes World Health Organization, 2018, p. 88.

[9] S. Edition, IDF Diabetes Atlas, seventh ed., 2015.

[10] International Diabetes Federation and Nam Han Cho (chair), et al., Eighth edition 2017, 2017.

[11] IDF, IDF Diabetes Atlas 2019, ninth ed., 2019.

[12] B. Davazdahemami, H.M. Zolbanin, D. Delen, An explanatory analytics frame- work for early detection of chronic risk factors in pandemics, Healthc. Anal. 2 (January) (2022) 100020, http://dx.doi.org/10.1016/j.health.2022.100020.

[13] M. Samieinasab, S.A. Torabzadeh, A. Behnam, A. Aghsami, F. Jolai, Meta-health stack: A new approach for breast cancer prediction, Healthc. Anal. 2 (October 2021) (2022) 100010.

[14] S.M. Ganie, M.B. Malik, T. Arif, Machine learning techniques for diagnosis of type 2 diabetes using lifestyle data, in: International Conference on Innovative Com- puting and Communications, in: Advances in Intelligent Systems and Computing, vol. 1394, Springer, Singapore, 2022, pp. 487–497.

[15] N. Nissa, S. Jamwal, S. Mohammad, Early detection of cardiovascular disease using machine learning techniques an experimental study, Int. J. Recent Technol. Eng. 9 (3) (2020) 635–641.

[16] F. Anwar, Qurat-Ul-Ain, M.Y. Ejaz, A. Mosavi, A comparative analysis on diagno- sis of diabetes mellitus using different approaches – A survey, Inf. Med. Unlocked 21 (April) (2020) 100482. S.M. Ganie, M.B. Malik, T. Arif, Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches, J. Diabetes Metab. Disord. 2022 (2022) 339–352.

[17] V. Rawat, Suryakant, A classification system for diabetic patients with machine learning techniques, Int. J. Math. Eng. Manag. Sci. 4 (3) (2019) 729–744,

[18] S.M. Ganie, M.B. Malik, T. Arif, Early prediction of diabetes mellitus using various artificial intelligence techniques: A technological review, Int. J. Bus. Intell. Syst. Eng. 1 (4) (2021) 1–22.

[19] S. Jamwal, S.M. Najmu Nissa, Heart disease prediction using machine learning, in: Lect. Notes Networks Syst., vol. 203 LNNS, (no. 67) 2021, pp. 653–665.

[20] V. Chang, V.R. Bhavani, A.Q. Xu, M. Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, Healthc. Anal. 2 (2021)

(2022) 100016, http://dx.doi.org/10.1016/j.health.2022.100016.

[21] Anaconda Inc, Anaconda distribution, Anaconda, 2019, [Online]. Available: https://www.anaconda.com/distribution/.

[22] S. Raschka, J. Patterson, C. Nolet, Machine learning in python: Main develop- ments and technology trends in data science, Mach. Learn., Artif. Intell., Inf. 11 (4) (2020).

[23] Jazayeri, O.S. Liang, C.C. Yang, Imputation of missing data in electronic health records based on patients' similarities, J. Healthc. Inform. Res. 4 (3) (2020) 295–307.

[24] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for type

[25] 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018).

[26] Y. Li, H. Li, H. Yao, Analysis and study of diabetes follow-up data using a data-mining-based approach in New Urban Area of Urumqi, Xinjiang, China, 2016–2017, Comput. Math. Methods Med. 2018 (2018) http://dx.doi.org/10. 1155/2018/7207151.

[27] M.K. Hasan, M.A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531.

[28] S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, Int. J. Cogn. Comput. Eng. 2 (November 2020) (2021) 40–46.

[29] S.M. Ganie, M.B. Malik, Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus, Int. J. Med. Eng. Inform., in press.

[30] M.B. Malik, S.M. Ganie, T. Arif, Machine learning techniques in healthcare informatics: Showcasing prediction of type 2 diabetes Mellitus disease using lifestyle data machine learning in healthcare, Predict. Model. Biomed. Data Min. Anal. (2022).

[31] M. Izadikhah, A fuzzy stochastic slacks-based data envelopment analysis model with application to healthcare efficiency, Healthc. Anal. 2 (February) (2022) 100038, http://dx.doi.org/10.1016/j.health.2022.100038.

[32] M.N. Algedawy, Detecting diabetes mellitus using machine learning ensemble, 670 Int. J. Comput. Syst. ISSN 03 (12) (2017) 670–677.

[33] P. Doupe, J. Faghmous, S. Basu, Machine learning for health services researchers, Value Heal. 22 (7) (2019) 808–815.

[34] S.K. Dehkordi, H. Sajedi, Prediction of disease based on prescription using data mining methods, Health Technol. (Berl) 9 (1) (2019) 37–44.